Integrative QTL analysis to identify causal candidate genes

Research rapport

Name: Melissa Spitteler
Student number: 970326790030
Supervisor: Harm Nijveen
Date: 10-07-2020

Summary

The engineering of plants traits is an imporant aspect of developing adaptive plants to biotic and abiotic stress. eQTLs are a powerfull tool in order to find variable regions in the genome and could therefor be usefull for the identifying of specific genes that contribute to plant traits. Currently the finding of causal genes in a QTL region is still a time-consuming process. Therefore this research will focus on developing a method to refine a candidate list of genes for eQTLs. Metabolic pathway information is used to extend the QTL analysis and to identify causal candidate genes for metabolic pathway enriched hotspot locations. Two methods are used, the identifying of *trans*-eQTL hotspot locations with pathway enrichment and the comparison of gene eQTLs with mQTLs to find overlapping QTL peaks from genes involved in the same pathways.

1. Introduction

Humanity will be facing immense population growth and food resource challenges in the near future. In order to catch up with the growing food demand, improving efficiency of harvest and developing adaptive crops to abiotic stress and biotic stress will be a priority. An important feature of crops is the variation in quantitative traits, this allows organisms to adapt to various environments (Baxter et al., 2010; Leinonen et al., 2013)

It is important to facilitate engineering of plants traits to create plants with the desired qualities (Lin et al., 2019). To achieve this goal, it is of importance to identify the specific genes that contribute to the traits. Quantitative trait locus (QTL) analysis has been developed to find regions on the genome which are associated with a particular trait. Currently the QTL approach is extended by using micro-arrays and RNA sequencing to measure expression of mRNA transcripts for genes. Differences in gene expression can be highly correlated with certain variable regions on the genome, thus identifying expression quantitative trait loci (eQTLs). Such regions may contain a single nucleotide polymorphism (SNP), large structural variants or copy number variants (Nodzak, 2020). Although expression quantitative trait loci is a powerful tool, the finding of causal genes underlying the expression QTLs is not straightforward (Bergelson & Roux, 2010; Cullingham et al., 2020). It is a time-consuming process to narrow down the list of candidate genes in a QTL region. The problem can be that there are either no obvious candidate genes or too many genes which could be potentially relevant for the specific trait (Nuzhdin et al., 1999).

For the model plant *Arabidopsis thaliana,* a number of genetical genomics studies uncovered many expression quantitative trait loci. With the use of markers throughout the whole genome, the eQTL peak with the highest LOD score is determined. The LOD score is a measure of the strength of evidence for the presence of a QTL at a particular location, the highest LOD score of eQTL peaks have the most evidence for the presence of the QTL. The success of QTL mapping is manly determined by the mapping resolution which depends on the size of the population and the number of recombination events (Loudet et al., 2002).

Two types of eQTLs are distinguished. If an eQTL arises on the location of the causal gene this is called *cis*-eQTL, a polymorphic site in the gene's own promoter that affects that gene's expression can cause this. A *trans*-eQTL is not located on the causal gene its location, but it occurs on the location of the gene's regulator and/or at a location of involvement in the same biochemical process (Fu et al., 2009; Vosman et al., 2019).This effect may arise when there is a polymorphism in the regulator (Doss et al., 2005). If many eQTLs are expressed on the same location, it is called a hot-spot, this can be the effect of a locus where a master regulator is located, several genes may show the same eQTL pattern due to this regulator (Hong & Breitling, 2008).

The platform AraQTL stores and combines the published *Arabidopsis* eQTL data and makes it possible to search this data in a simple way. Large data sets with genetic interactions are hard to tackle for most biologists. AraQTL provides an ease of access and possibilities for interactive exploration. This facilitates the generating of new hypothesis, follow-up laboratory experiments and enables the re-use of previously published data.

To summerize, the engineering of plants traits is an imporant aspect of developing adaptive plants to biotic and abiotic stress. eQTLs are a powerfull tool in order to find variable regions in the genome and could therefor be usefull for the identifying of specific genes that contribute to plant traits. Currently the finding of causal genes in a QTL region is still a time-consuming process. Therefore this research will focus on developing a method to refine a candidate list of genes for eQTLs.

The project will include two methods. Both methods will incorporate expression quantitative trait loci and metabolic pathway information. Metabolic pathway information in *A. thaliana* is well researched, this makes it possible to use this extensive data to explain expression QTLs. The metabolic pathways also have an important role in many systems which may influence important traits for crops. The first method is to locate hotspots locations of *trans*-eQTLs and test for metabolic pathway enrichment. The *trans*-eQTL hotspot location can be caused by a regulator(s), since it is expected that this regulator regulates genes involved in similar processes, the metabolic pathway enrichment is performed. If an metabolic pathway enrichment is found, the genes involved in this pathway are candidates of the *trans*-eQTL hotspot and further analysis can be performed to find the regulator(s).

The second method is to compare eQTLs of individual genes with mQTLs. mQTLs have the same principle as eQTLs except that they are not from an individual gene but from a metabolism. The mQTLs and eQTLs are compared with the goal to find overlapping peak patterns of genes involved in the same metabolic pathway. If such overlap exists between an individual genes involved in the same metabolic pathway it is a candidate causal genes.

The found causal genes of eQTLs can be used to facilitate plant engineering for specific traits. The desired outcome is a method to optimally combine genomics data with prior knowledge on metabolic pathways in order to refine a list of causal genes for eQTLs.

2. Methods

2.1 *Trans*-eQTL hotspot locations

2.1.1 Data collection and peak calling

Data is collected from the paper Serin et al., 2018. The dataset consists of RNA-seq data of an *Arabidopsis* Bay-0 × Shahdara RIL Population of dried seed, a QTL analysis is performed on the RNA-seq data. With the QTL analysis data, a QTL peak caller script written by Linda van Bemmelen in python language is used to call all the eQTL peaks. The QTL peak caller only calls one peak per chromosome, the LOD threshold is set to 3.0. The parameters absolute drop and fraction drop are optional and are set to the recommended 1.5 and 0.10 (Linda van Bemmelen, 2018). These parameters define the peak boundaries. The absolute LOD drop is subtracted from the peak LOD score and the nearest markers that have a LOD score below this threshold are taken as boundary markers. The fraction drop is a fraction which is multiplied by the peak LOD score used as the LOD drop.

2.1.2 Semantic web

The semantic web is a pre-existing environment and used to collect relevant data. The metabolic pathway information is available in a semantic web (Tang, 2019.). This semantic web contains data from StringDB, KEGG and GO databases. In order to collect this data, queries are made in the SPARQL language. First, all genes and their locations of *Arabidopsis Thaliana* are retrieved from the semantic web.

2.1.3 Filtering *cis/trans*-eQTLs

The gene locations and QTL peaks are used in a python script that divides all called QTL peaks into *cis*-QTLs and *trans*-QTLs, depending on the location of the gene and the QTL. If the gene location is inside the QTL it is categorised as a *cis*-eQTL, if the gene location is outside the QTL it is categorised as a Tans-eQTL. *Trans*-eQTLs are collected in a python dictionary with the peaking marker as key and the genes peaking at this marker as value.

2.1.4 Metabolic pathway information

From the semantic web all KEGG metabolic pathways linked to *Arabidopsis* genes are retrieved. Also included are the metabolic pathways in *Arabidopsis* available on Aracyc, this database is not present in the semantic web so it is chosen to download the data and use python to link this data. By including both the KEGG database as the Aracyc database, most available genes linked to metabolic pathways are included. With another python script all *trans*-eQTL genes are linked to all pathways the genes are involved in. Again, a dictionary was composed of peaking marker locations as keys and pathways as value.

2.1.5 Metabolic pathway enrichment

In order to find *trans*-eQTL hotspot locations with an enrichment of a metabolic pathway a Fisher's exact test is conducted. The Fisher's exact test takes four parameters, these are shown in Table I.
To conduct the Fisher's exact test the python library scipy is used with the function stats.fisher_exact. This returns the odds ratio and p-value of all tested pathways. To control the false discovery rate the Benjamini and Hochberg correction is conducted with the python3 library statsmodels.sandbox.stats.multicomp with multipletests. Results with a p-value of below 0.05 and an odds ratio of above 1 are collected.

Table I, Layout of the Fisher's exact test input

	#genes with pathway	#genes without pathway
#genes in set	a	b
#genes not in set	c	d

All hotspot location results are manually inspected, this includes the metabolic pathways and all genes causing the *trans*-eQTL hotspot location. This is done to select the most promising hotspot locations, which is based on the amount of genes found and the function of these genes. This is also done to find an explanation for the hotspot *trans*-eQTL location and refine a candidate list. Additionally, from the genes expressing a QTL on the hotspot location, all interacting genes are retrieved from the semantic web to find potential new genes involved in the metabolic pathway.

Furthermore, the previously described workflow is also used on multiple other datasets. The datasets are collected from AraQTL and included (Cubillos et al., 2012),(Joosen et al., 2012), (Lowry et al., 2013) and (Snoek et al., 2013).

2.2 Correlating metabolic QTLs

Metabolic expression data is collected from Serin et al., 2017. The metabolic data originates from metabolic profiling of dry harvested seeds of the RIL population by GC-TOF-MS. The individual gene expression data is collected as written in section '2.1.1 Data collection and peak calling'.

2.2.1 Clustering

With the expression data, the spearman correlation is calculated pairwise for the metabolites in a python script, with all the pairwise correlation data a hierarchical clustering is performed. To visualise the data, a dendrogram is made in python with the toolboxscipy.cluster.hierarchy.dendrogram. In the dendrogram, the length of the two legs of the U-link represent the distance between the metabolite clusters. Metabolic pairs with a high correlation have a low distance in the dendrogram. Based on the distance of metabolites in the dendrogram, the chosen distance is 6.5 and the metabolites are divided into ten groups. This distance is based on the amount of groups formed, amount of metabolites in each cluster. These groups are expected to have similar expression patterns and therefor may share QTLs caused by the same genes.

2.2.2 Peak calling

Next, the QTL peaks are called using a QTL peak caller script, written by Linda van Bemmelen in python language. The script searches for the marker with the highest LOD score per chromosome and based on the parameters absolute LOD drop and fraction drop searches for the peak boundaries. The LOD threshold is set to 3.0. The parameters absolute drop and fraction drop are optional and are set to the recommended 1.5 and 0.10. The same approach is used for the individual gene expression data.

For each peak found, the boundary markers of the peaks with LOD above 3 are collect. The location in between the boundary markers of the mQTL peaks are compared to QTL peaks of the individual genes. If a gene has a peaking QLT marker inside the boundary markers of the mQTL, the gene is collected.

2.2.3 Pathway information

For each gene causing an eQTL that is overlapping with an mQTL, all pathways that the gene is in involved are obtained from the KEGG database and the Aracyc database. For each of the ten groups all data is collected in excel. By selecting a specific metabolic pathway in excel, it is shown if multiple genes are found for this metabolic pathway and if these genes are located on different chromosomes.

Finally, with the previously described method a list of potential genes involved in mQTL can be refined.

2 Results

3.1 *Trans*-eQTL hotspots

To start the project, eQTL data was collected from E. Serin, 2018. This paper describes the construction of a high-density genetic map from RNA-Seq Data for an *Arabidopsis* Bay-0 × Shahdara RIL Population. The high-density map is constructed with 1059 markers and includes 27,402 *Arabidopsis* genes[1]. Next, all peaks are identified with a LOD score above 3.0. Over 17,000 peaks were identified including 12,217 genes.

An eQTL located on the same location as the gene itself is a *cis*-eQTL and a QTL not located on the same location as the gene is called a *trans*-QTL. For the purpose of identifying hotspot locations of

[1] https://plants.ensembl.org/Arabidopsis_thaliana/Info/Annotation/, consulted on 28-06-2020

trans-eQTL, the called *trans*-QTL peaks were filtered. This resulted in 8766 *trans*-eQTL.

For each QTL the marker with the highest LOD score is determined, all the genes with the same peaking location are grouped. The collected genes with the same QTL peak are linked to metabolic pathway information available from the KEGG database. This facilitated the pathway enrichment analysis by the Fisher's exact test. If an enriched metabolic pathway is found on a hotspot location, the genes found can be potential candidates causing the hotspot eQTL location.

Fisher's exact test did not yield any pathway enriched hot-spot locations. Therefor the choice was made to not only include genes with the same peaking marker but also include both neighbouring markers of the peaking marker. This broadened the peaking location of the QTL and thus the number of genes included. Again, a Fisher exact test was performed in order to find enriched metabolic pathway hotspot locations, this resulted in 10 *trans*-eQTL pathway enriched locations identified. In Figure 1 the 5 five *Arabidopsis* chromosomes are shown with all the markers located on the hotspot locations. Each hotspot location is shown as at least two markers (the low genome position neighbour marker and the high genome position neighbour marker). In some cases, the hotspot has an even broader range. This is the case when one of the neighbour markers is also enriched with the metabolic pathway and their neighbours are included as well.

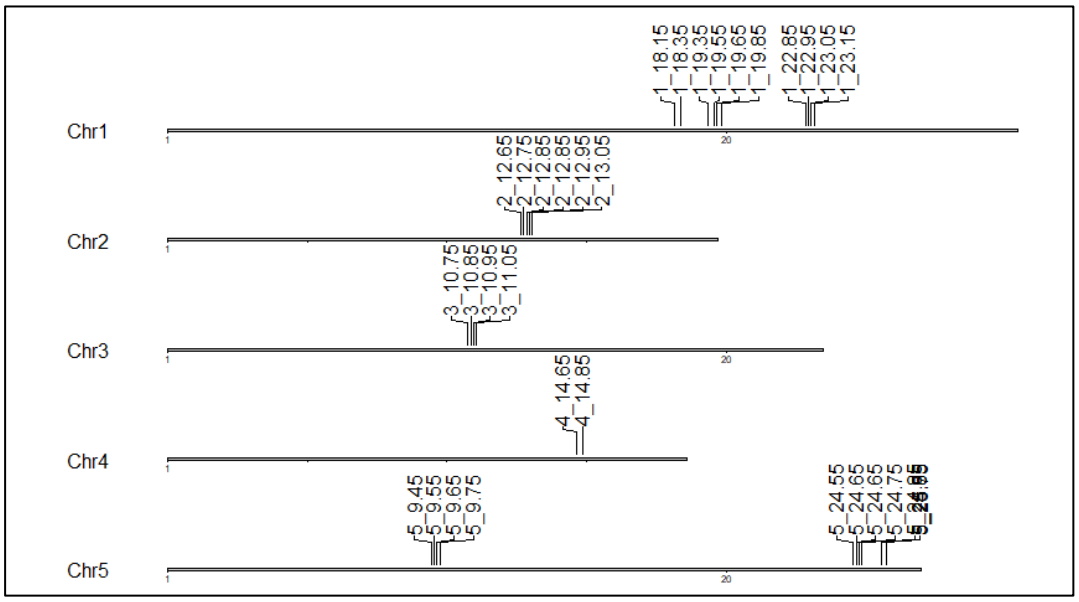


Figure 1, The five Arabidopsis chromosomes, with all the markers located on enriched pathway hotspot locations. From left to right: chromosome 1: marker 1_18.15 – marker 1.18_35: Map03060 = Protein export, marker 1_19.35 – marker 1_19.85: Map00903 = Limonene and pinene degradation, marker 1_22.85 – marker 1_23.15 Map00905 = Brassinosteroid biosynthesis. Chromosome 2: marker 2_12.65 – marker 2_13.05: Map03010 = Translation ribosomal. Chromosome 4: marker 4_14.65 – marker 4_14.85: map03022 = Basal transcription factors. Chromosome 5: marker 5_9.45- marker 5_9.75: Map03015 = mRNA surveillance pathway, marker 5_24.65 – marker 5_24.85: Map00970 = Aminoacyl-tRNA biosynthesis, marker 5_24.65 – marker 5_24.85: Map03008 = Ribosome biogenesis in eukaryotes, marker 5_25.55 – marker 5_23.35:Map03040 = Spliceosome.

Table II, enriched metabolic pathway hotspots found with amount of genes found on the hotspot location involved in the enriched metabolic pathway

Chromo-some	Markers	Metabolic pathway	Number of genes found
1	marker 1_18.15 – marker 1.18_35	Map03060 = Protein export	3
1	marker 1_19.35 – marker 1_19.85	Map00903 = Limonene and pinene degradation	2
1	marker 1_22.85 – marker 1_23.15	Map00902 = Monoterpenoid biosynthesis	2
2	marker 2_12.65 – marker 2_13.05	Map03010 = Translation ribosomal	83
4	marker 4_14.65 – marker 4_14.85	Map03022 = Basal transcription factors	2
5	marker 5_9.45 - marker 5_9.75	Map03015 = mRNA surveillance pathway	5
5	marker 5_24.65 – marker 5_24.85	Map00970 = Aminoacyl-tRNA biosynthesis	3
5	marker 5_24.65 – marker 5_24.85	Map03008 = Ribosome biogenesis in eukaryotes	7
5	marker 5_25.55 – marker 5_23.35	Map03040 = Spliceosome	4

In Table II the enriched metabolic pathway locations from Figure 1 are shown. A more detailed table of results is shown in Appendix I. Each hotspot location was investigated individually in order to explain the QTL expressed on these locations. Each gene found at the hotspot location with the enriched pathway is checked to see if they have a clear link with the found metabolic pathway and with each other. Although all pathway locations were significantly enriched found, most of the hotspot locations did not have enough genes included to explain the QTL. When a *trans*-eQTL hotspot is found, it is assumed that there is a regulator involved to regulate the genes expressed at the hotspot location. When only a couple genes are found at a metabolic enriched hotspot location it becomes difficult to find the regulator of the genes. The only hotspot location with enough genes to further analyse is chromosome 2 marker 2_12.65 until marker 2_13.05.

With the use of AraQTL the marker locations (12650000:13050000) on chromosome 2 are visualized with the 83 genes that were found metabolic pathway enriched at this location. As shown in Figure 2, the 83 genes all have similar expression patterns, especially on chromosome 2. Since the QTLs on chromosome 2 are all *trans*-QTL, it is assumed that one or more regulators are responsible for the expression of these genes on the pathway enriched hotspot location.

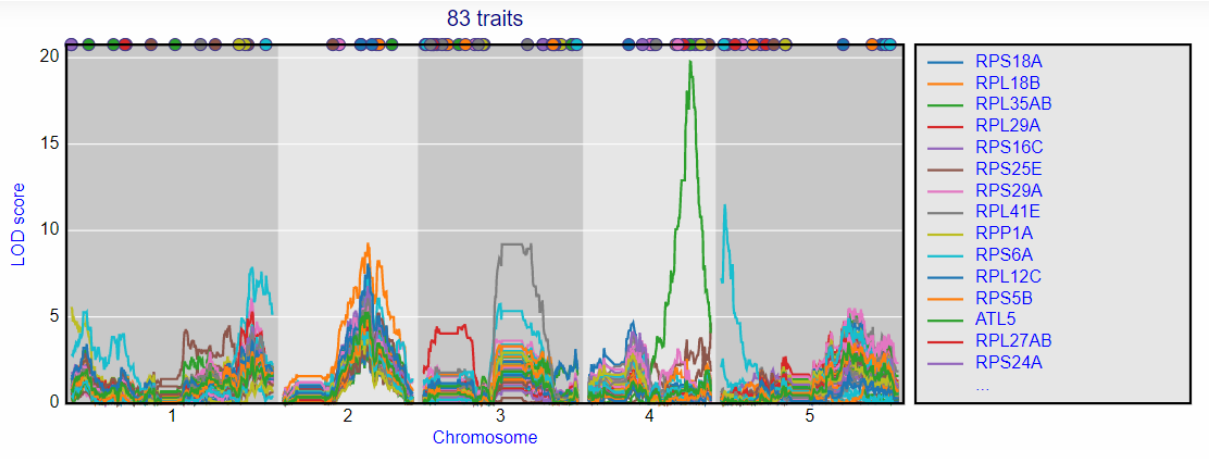


Figure 2, visualisation in AraQTL of the 83 genes found with pathway translation ribosomal

In order to validate this hypothesis, the program TF2Network is used. This program predicts potential transcription factor regulators for a set of functionally related genes in *Arabidopsis thaliana*. It uses TF binding site information and integrates co-expression data, experimental Protein-DNA and Protein-Protein Interactions to make this prediction. Three possible regulators were found, AT5G67580 with 38 hits (q-value of 5.44e-10). AT1G72740 37 hits (q-value of 1.52e-9). AT3G49850 32 hits (2.68e-6). All three regulators were checked, but none of the genes have the similar peaking markers as the previous found 83 genes. Only one of the regulators AT5G67580 (TRB2) shows a peak of expression in the same area of marker 2_12.65 until marker 2_13.05, but the peak of expression is at marker 2_13_65 with a LOD score of -4.86. In linkage analysis a negative LOD score is a sufficient indication that a given variant is not linked with the trait gene (Ott et al., 2015).

Although no gene was found that could be responsible for the hotspot location. The interest occurred to expand the metabolic pathway with the genes that were found. Metabolic pathway information can expand the knowledge to understand how an organism works. With use of the semantic web, from the STRING database all interacting genes of the 83 found genes were extracted. These interactions are derived from high-throughput experimental data, from the mining of databases, literature, and from predictions based on genomic context analysis (von Mering et al., 2005). A graph was constructed (Figure 3) with Cytoscape of all 83 genes found at the hotspot-location (blue) and the 52 interacting genes are oval shaped(green/red). This was done for the discovery of potential genes involved in the same metabolic pathway. Of the found genes, it was checked if they were already involved in the pathway (translation ribosomal) or not. In the graph, all the red genes with an oval shape, where not already included in the metabolic pathway.

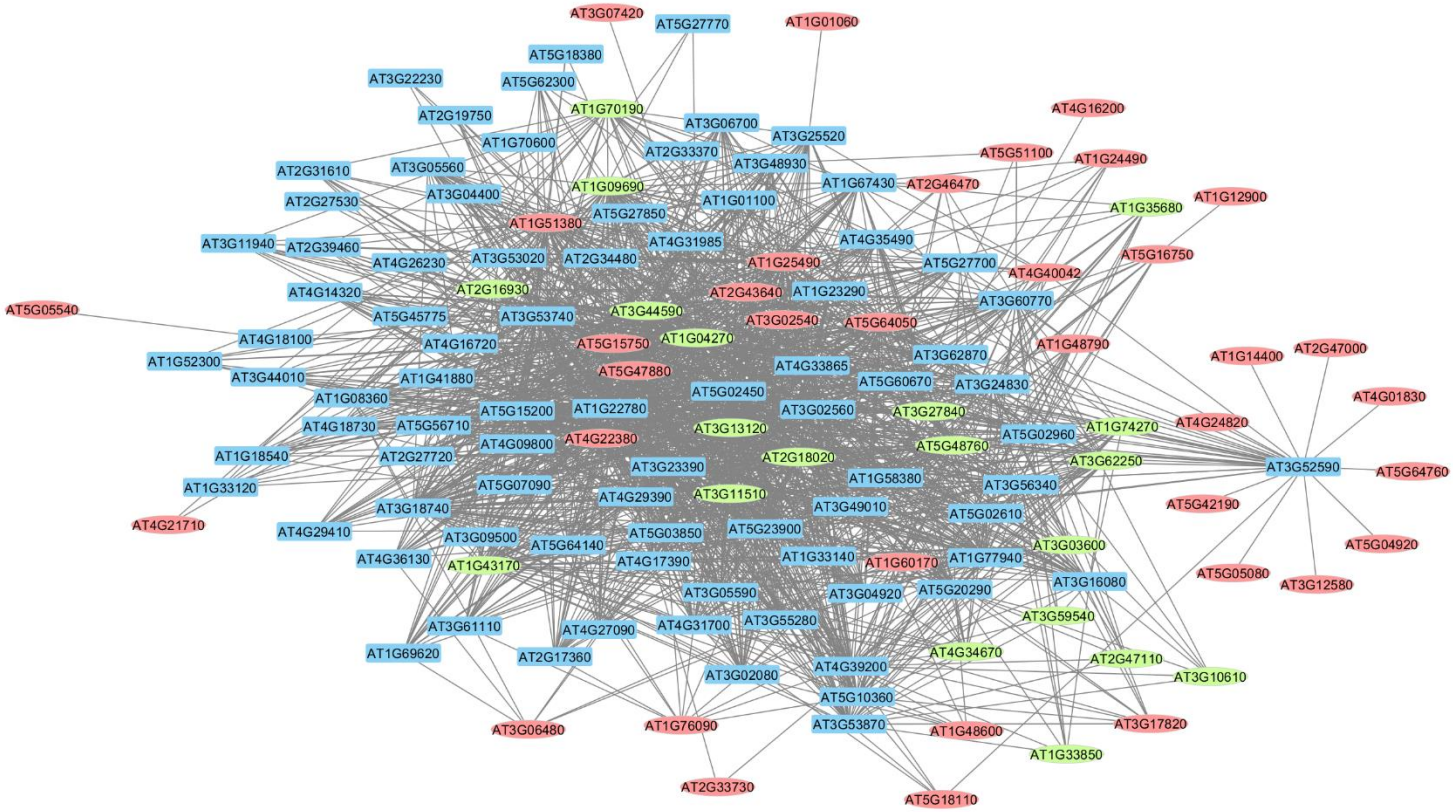


Figure 3, Cytoscape network of the 83 genes found with pathway translation ribosomal (blue) and all their interacting genes (green/red and oval shaped). The green genes are also involved in the pathway translation ribosomal. The red genes are not involved in the pathway translation ribosomal.

Table III, genes found interacting with genes in the pathway translation ribosomal.

AT1G48790	AMSH-like ubiquitin thioesterase 1
AT3G02540	Encodes a member of the RADIATION SENSITIVE23 (RAD23) family: AT1G16190(RAD23A), AT1G79650(RAD23B), AT3G02540(RAD23C), AT5G38470(RAD23D). RAD23 proteins play an essential role in the cell cycle, morphology, and fertility of plants through their delivery of UPS (ubiquitin/26S proteasome system) substrates to the 26S proteasome
AT3G02540	Glutamate-tRNA ligase. Targeted to mitochondria and chloroplast. Its inactivation causes developmental arrest of chloroplasts and mitochondria in Nicotiana benthamiana.
AT5G47880	Encodes a eukaryotic release factor 1 homolog. Cosuppression of the gene's expression results affects cell elongation of the inflorescence stem, specifically the internodes, and radial cell division. Expression of the protein is primarily observed in the vascular system and in actively growing and elongating zones.
AT1G51380	DEA(D/H)-box RNA helicase family protein
AT1G25490	One of three genes encoding phosphoprotein phosphatase 2A regulatory subunit A; Recessive ethylene-response mutant EER1 displays increased ethylene sensitivity in the hypocotyl and stem
AT4G22380	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT5G15750	Alpha-L RNA-binding motif/Ribosomal protein S4 family protein
AT2G43640	Signal recognition particle, SRP9/SRP14 subunit

The organic layout used for the network gives a visualization of highly connected backbone regions in the network, in this case the genes located in the core of the network[2]. The genes located in the core of the network have the most interaction with the other genes and are therefore most likely to be involved in the metabolic pathway 'translation ribosomal'. As shown in table III, some of the genes encode a protein which is involved in the ribosomal translation. Other genes are involved in different components of the cell than the ribosomes, such as the chloroplasts and mitochondria. To prove these genes are also involved in the ribosomal translation, experimental testing is needed.

To continue the project and to find more enriched pathways on *trans*-eQTL hotspot locations, the database PMN Aracyc was added. This database contains 517 metabolic pathways in *Arabidopsis*. With the addition of the extra database, the pipeline was run on the Serin *et al,.* 2018 data. This did not yield any new enriched metabolic pathway *trans*-eQTL hotspots locations.

Additionally, the pipeline is tested on four extra datasets (Snoek *et al,.* 2012, Cubillos *et al,.* 2012, Lowry *et al,.* 2013, Joossen *et al,.* 2012). All these datasets use different *Arabidopsis* populations, Snoek *et al,.* 2012 uses a Ler x Cvi population in Rosette stage with samples taken from leaves 24 days after germination. Cubillos *et al,.* 2012 uses a Cvi x Col population, also in the Rosette developmental stage. Lowry *et al,.* 2013 uses a Tsu x Kas population in the Rosette stage and Joosen *et al,.* 2012 uses a Bay x Sha RIL population from seed. In all these different populations, some similar eQTL can be expected but also some different ones.

Although all of the datasets have many significant results, none of the results have a odds ratio of above 1. The odds ratio is a measure that shows how strong the association is, if the odds ratio is below 1 it means that there is no association between genes found and the metabolic pathway.
In view of the number of genes found at a specific marker with a pathway, generally there are only 1 or 2 genes found. That being the case, there is a lack of pathway enrichments in these datasets.

[2] https://yed.yworks.com/support/manual/layout_smartorganic.html consulted on 10-07-2020

Alternatively, we choose to reduce the required LOD from 3.0 to 1.5. Since the QTL peak caller only calls one peak per chromosome, this will fill the gaps of chromosome without a peak called. The pipeline was run again on the five datasets. Again, this did not yield any results.

The choice is made to use the peak boundary markers of the QTL peaks in order to collect more genes. The metabolic pathways are linked to the genes and the Fisher exact test is performed. This resulted in multiple metabolic pathway enriched locations. The amount of metabolic enriched *trans*-eQTL hotspot differ per dataset used. For Cubillos 2012 only 29 metabolic pathway enriched hotspots were found, whereas Joossen 2012 yielded 743 metabolic pathway enriched hotspots. Theses metabolic pathway enriched hotspots locations need manual reviewing, although the p-value is significant and odds ratio is above 1. The results are not always convincing, when for example at the analysis of Joosen 2012, between markers RSM_1_12.95 until RSM_1_16.95 (chromosome 1, 12950000 : 16950000) four genes are found with pathway PWY-3781 (aerobic respiration I (cytochrome c)) out of 289 genes involved in this pathway. The boundary markers of this peak are quite large, to only find 4 genes involved in the specific pathway is not a convincing result. In the dataset of Cubilos 2012 (Appendix II) tare some interesting results, for example between marker c4_00641 until c4_04877 (chromosome 4, 641363: 4877120) , 12 genes are found involved in PWY-5143 (long-chain fatty acid activation) out of 12 genes known to be involved in this pathway.

However, because the widest location of the QTL peaks are used the enriched found pathways are not specifically on a hotspot location but in a wider range, depending on the boundary markers of the peak. Regardless of this downside, if a high percentage of genes of a specific pathway can be found with this method at a location, it can still provide a list of potential causal genes of the pathway enriched hotspot location.

3.2 Metabolic QTLs

For the second part of the project the aim is to find causal genes involved in mQTLs. This is done by clustering correlating mQTLs and comparing QTL peaks to individual gene QTL peaks. With the use of metabolic pathway information, genes involved in the same pathways with a similar LOD profile of the mQTLs can be matched.

Clustering is performed to make groups of mQTLs with a similar LOD expression profile. This is done to be able to find multiple similar results in one group of correlation mQTLs. Some of the metabolites are identified, this identification is based on spectral similarities and comparing the retention indices calculated (E. Serin, 2018). The metabolites are small molecules such as sucres or amino acids. Based on the clustering of the mQTLs, the mQTLs are divided into ten groups. For each group all QTL peaks are called and all peaks are called of the individual genes. The peaking markers of each mQTL peak are matched with the peaking markers of the individual gene QTL peak from the dataset Serin 2018. Metabolic pathway information is collected for each gene, if multiple genes have overlapping QTL peaks as the mQTLs and have common pathways they are further analysed. Because multiple interesting results are found with this method, one example is taken. In table IV group 1 is shown, these mQTLs have a distance below in the dendrogram 6.5 in the clustering.

Table IV, group 1 based on spearman clustering with a distance of dissimilarity in the dendrogram of below 6.5.

RI_1551.377948
Sorbitol.galactitol
Galactaric.acid
Malonic.acid
MPIMP.ID.144003.21.1
RI_2046.708054

Salicilyc.acid
Ferulic.acid
RI_2087.650937

With the matching of QTL peaks of group one with de individual genes, 11 pathways are found with more than five genes with similar QTL peaks per pathway. By manually going through these results one pathway shows to be the most interesting. This is decided by the number of genes matching QTL peaks and the distribution of these genes over the five *Arabidopsis* chromosomes. The most interesting metabolic pathway is map04075: Plant hormone signal transduction (KEGG). There are two metabolites involved in this finding, RI_1551.377948 and Malonic acid.

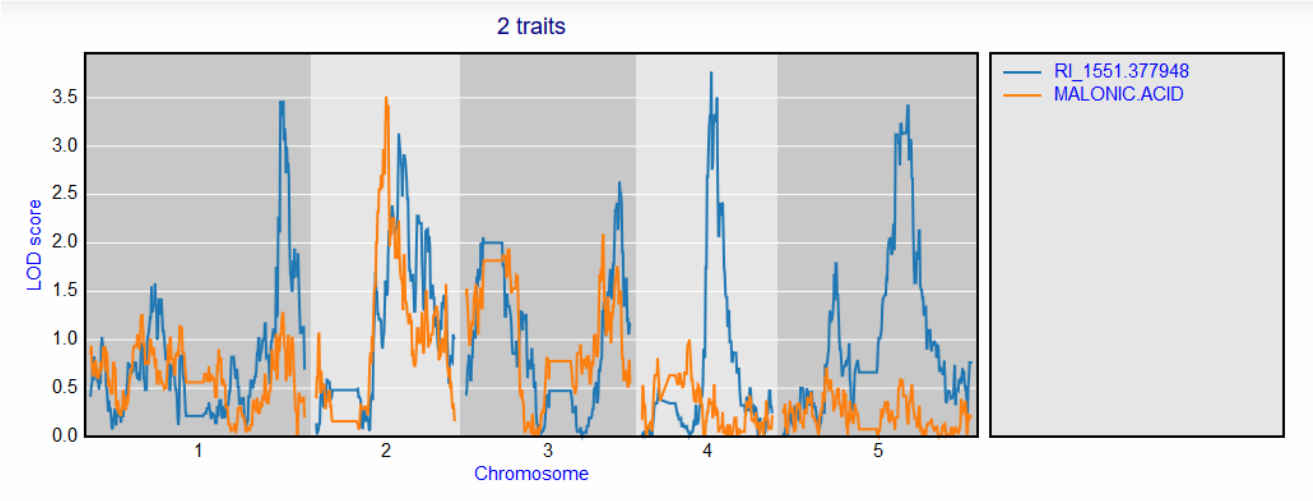


Figure 4, LOD profile for metabolites RI_1551.377948 and Malonic acid visualised in AraQTL.

As shown in the Figure 4, peaks with a LOD above 3 are mainly due to mQTL RI_1551.377948. The metabolite Malonic acid only contributes one peak at chromosome 2. Also visible are the similarities in LOD profile on chromosome 2 and 3. In Figure 5 the network genes with overlapping QTL peaks are shown.

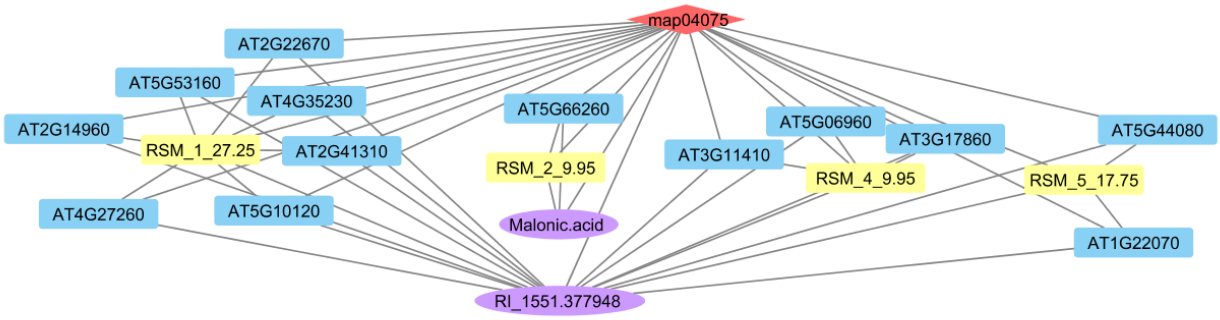


Figure 5, network of the found genes(blue) linked to the peaking markers (yellow) of QTLs originated from RI_1551.377948 and Malonic acid (purple).

To verify the result, the LOD score profiles of the genes is also visualized with the use of AraQTL. This result is shown in Figure 6. Again, the expression profile is extremely similar for these genes, with only three not overlapping peaks at chromosome 1 and 5. This verifies that these genes are regulated in a similar matter and make them suitable candidates for the found peaks of the metabolites. In appendix III, the specific description for all the genes is shown. The pathway involves multiple reactions (Appendix IV), that included Cell enlargement/plant growth, cell division/shoot initiation, stem growth/induced germination, stomatal closure/seed dormancy, fruit ripening/senescence,

senescence/stress response and disease resistance. From the two metabolites only one is identified as a secondary metabolism, malonic acid. Malonic acid itself is an organic acid and play a role if different pathway such as Fatty acid Biosynthesis, Pyrimidine metabolism and beta-Alanine metabolism (Chen et al., 2011). It is also a precursor important enzyme Malonyl-CoA (Gueguen et al., 2000). Malonyl-CoA has been shown to be essential for growth and development of *Arabidopsis*. The enzyme is a precursor for fatty acid synthesis and elongation and therefor one of the building block of many organic compounds (Chen et al., 2011). This supports the found link between Malonic acid and the pathway 'Plant hormone signal transduction', where a lot of growth-related reactions are taking place. It also makes it likely that the other metabolite RI_1551.377948 also plays a role in growth related reaction, although it is impossible to pinpoint the exact function.

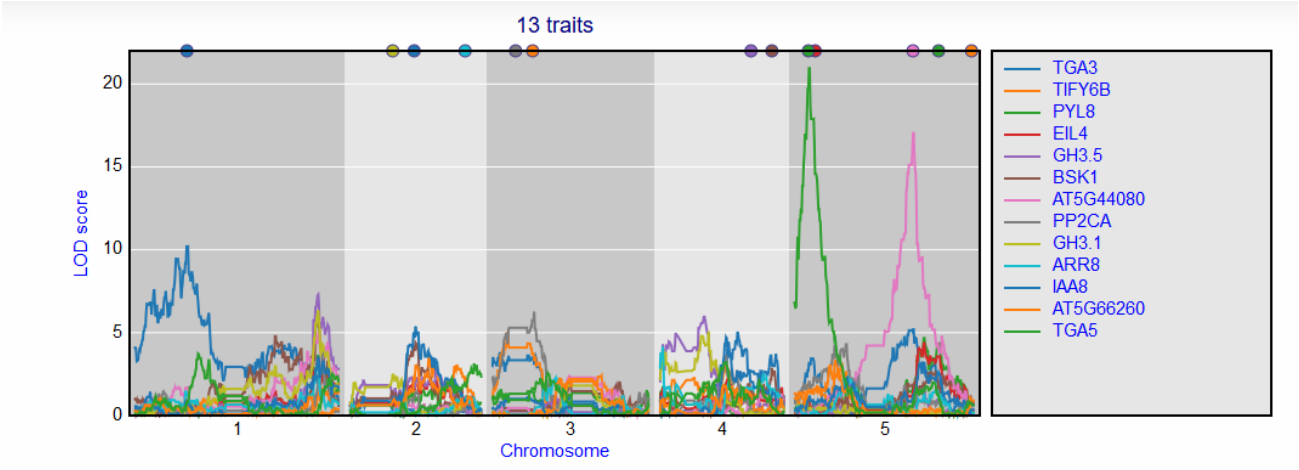


Figure 6, LOD profile of overlapping genes to metabolites RI_1551.377948 and Malonic acid.

The results of the other mQTL groups are various (Supplementary data I). Some of the groups, have too many possible pathways found where the distribution of genes for the chromosomes is evenly throughout all these pathways. Moreover, multiple groups have QTL peaks linked to genes located on only two different chromosomes. This makes it impossible to pinpoint the correct pathways and thus the correct causal candidate genes. However, there is also a straightforward results, Group 5 (Malic acid, MPIMP.ID.183003.21, N.Acetylglutamic acid, RI_2860.615732) only has two pathways found map01100 and PWY-3781, map001100 is the overview of all metabolic pathways and therefor found at every group. The other pathway found is aerobic respiration I, with 6 genes found (AT1G26900, AT1G70760, AT3G02090 , AT4G14050 , AT5G08490, AT4G39952) on chromosome 1, 4 and 5.

3 Discussion

The aim of this research was to use metabolic pathway information to extend the QTL analysis and to identify causal candidate genes for metabolic pathway enriched hotspot locations. Two methods were used, the identifying of *trans*-eQTL hotspot locations with pathway enrichment and the comparison of gene eQTLs with mQTLs to find overlapping QTL peaks from genes involved in the same pathways.

With the first method, the *trans*-eQTL hotspot location it was expected to yield multiple metabolic pathway enriched *trans*-eQTL hotspots. This was expected because there were a lot of *trans*-eQTLs and hotspot locations. However, this was not the case. Although many QTL peaks can be located on the same location of the genome, the genes involved in this QTL peaks also need to be involved in the same pathways in order to find enrichment. The first bottleneck is the number of genes linked in a pathway. First, only the KEGG database was used in the pipeline. With the addition of the Aracyc database, it was expected to yield more eQTL hotspots. In contrast to what was expected, no extra eQTL hotspots where found. The Aracyc database consists of 576 *Arabidopsis* pathways whereas the KEGG database consists of 138 pathways. It can be assumed that in each Aracyc pathway there will be less genes involved than in a KEGG pathway. Most of the Aracyc pathways consist of 5 to 20 genes each, whereas the KEGG pathways consist of 30 to 50 genes per pathway, excluding the KEGG overview pathways. The likelihood of finding a couple genes in a enriched KEGG pathway is higher. Quite some marker locations found with an enrichment for a specific pathway with a significant result. But because a filtering of results with odds ratio above 1 is done, none of the significant results satisfy this filtering. An odds ratio below 1 means that there is a lower odd of association between the number of genes found and the hotspot location. This is probably due to the low number of genes found at a pathway, e.g. 1 or 2 genes. That being the case, there is a lack of pathway enrichments in these datasets.

In order to find more results, it was chosen to lower de LOD score threshold to 1.5. This could cause QTL peaks called which might be noise. Although in the QTL peak caller used this should not be a great issue, since only one peak per chromosome is called. The highest peak will always be included for each chromosome. Decreasing the LOD-value from 3.0 to 1.5 will not give drastic changes in the called peaks, it will fill in gaps of chromosomes that did not already have a QTL peak.

Another approach was to take the boundary location of the QTL peak above LOD threshold. This did yield pathway enriched locations but also make the results less reliable. By taking a larger area, that of the boundary location of the QTL peak. The number of genes having a QTL peak in that area will also increase. This makes the likelihood of finding multiple genes with the same pathway larger. It does give a candidate list of genes that could be causal for the QTL peak. In order to check if this method is indeed less reliable, it could be tested on random pieces of chromosomes with multiple iterations to see if metabolic pathway enrichment is found on random locations with these larger boundary locations than if only the peaking marker and its neighbour markers are included.

For the second method, some of the correlation groups gave good results. Although the method can be further improved. Currently, the end result should still be manually assessed in order to find the best fitting results. Especially when multiple pathways are found, it is hard to pinpoint the pathway, also because multiple pathways could be involved. Another possible improvement is the incorporation of QTLs with a negative LOD score. Genes causing a negative LOD score could also be involved in metabolic pathways, it is expected that a negative effect can influence a positive reaction.

Another addition could be to not only compare mQTLs by individual peak QTL, but incorporate the whole LOD profile. Especially when LOD profiles show a lot of similarity this could help to find the correct genes involved in mQTLs.

As described before, both methods can be further improved but need more testing in order to find reliable results. Overall, the methods are a good starting point for further development.

4 Conclusion

The *trans*-eQTL method showed multiple *trans*-eQTL hotspot enriched pathway locations. Of these locations, the hotspot location enriched for the pathway 'ribosomal translation' was further analysed. 83 genes were found with a QTL peak on hotspot location and 52 genes where identified to be interacting. 37 genes were found as potentially to be involved in the pathway ribosomal translation.

The metabolites were divided into ten groups, based on spearman correlation of their LOD score profiles and the maximum distance of 6.5 in the dendrogram. For the first group, the pathway map04075: Plant hormone signal transduction was found, based on two mQTLs (RI_1551.377948 and Malonic acid). The other result is for group 5 (Malic acid, MPIMP.ID.183003.21, N.Acetylglutamic acid, RI_2860.615732), were the pathway aerobic respiration I was found, with 6 genes found (AT1G26900, AT1G70760, AT3G02090 , AT4G14050 , AT5G08490, AT4G39952) on chromosome 1, 4 and 5.

Both methods did result in some potential candidate genes for QTLs.

References

Baxter, I., Brazelton, J. N., Yu, D., Huang, Y. S., Lahner, B., Yakubova, E., Li, Y., Bergelson, J., Borevitz, J. O., Nordborg, M., Vitek, O., & Salt, D. E. (2010). A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1;1. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1001193

Bergelson, J., & Roux, F. (2010). *Adaptive walk Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana*. https://doi.org/10.1038/nrg2896

Chen, H., Kim, H. U., Weng, H., & Browse, J. (2011). Malonyl-CAa synthetase, encoded by ACYL ACTIVATING ENZYME13, is essential for growth and development of Arabidopsis. *Plant Cell*, *23*(6), 2247–2262. https://doi.org/10.1105/tpc.111.086140

Consultant, H. N. (2018). *eQTL Peak Calling in Python*.

Cubillos, F. A., Coustham, V., & Loudet, O. (2012). Lessons from eQTL mapping studies: Non-coding regions and their role behind natural phenotypic variation in plants. In *Current Opinion in Plant Biology*. https://doi.org/10.1016/j.pbi.2012.01.005

Cullingham, C. I., Peery, R. M., Fortier, C. E., Mahon, E. L., Cooke, J. E. K., & Coltman, D. W. (2020). Linking genotype to phenotype to identify genetic variation relating to host susceptibility in the mountain pine beetle system. *Evolutionary Applications*, *13*(1), 48–61. https://doi.org/10.1111/eva.12773

Doss, S., Schadt, E. E., Drake, T. A., & Lusis, A. J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research*. https://doi.org/10.1101/gr.3216905

Fu, J., Keurentjes, J. J. B., Bouwmeester, H., America, T., Verstappen, F. W. A., Ward, J. L., Beale, M. H., De Vos, R. C. H., Dijkstra, M., Scheltema, R. A., Johannes, F., Koornneef, M., Vreugdenhil, D., Breitling, R., & Jansen, R. C. (2009). System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nature Genetics*. https://doi.org/10.1038/ng.308

Gueguen, V., Macherel, D., Jaquinod, M., Douce, R., & Bourguignon, J. (2000). Fatty acid and lipoic acid biosynthesis in higher plant mitochondria. *Journal of Biological Chemistry*, *275*(7), 5016–5025. https://doi.org/10.1074/jbc.275.7.5016

Hong, F., & Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btm620

Joosen, R. V. L., Arends, D., Willems, L. A. J., Ligterink, W., Jansen, R. C., & Hilhorst, H. W. M. (2012). Visualizing the genetic landscape of Arabidopsis seed performance. *Plant Physiology*. https://doi.org/10.1104/pp.111.186676

Leinonen, P. H., Remington, D. L., Leppälä, J., & Savolainen, O. (2013). Genetic basis of local adaptation and flowering time variation in Arabidopsis lyrata. *Molecular Ecology*. https://doi.org/10.1111/j.1365-294X.2012.05678.x

Lin, F., Fan, J., & Rhee, S. Y. (2019). QTG-Finder: A Machine-Learning Based Algorithm To Prioritize Causal Genes of Quantitative Trait Loci in Arabidopsis and Rice. *G3 (Bethesda, Md.)*. https://doi.org/10.1534/g3.119.400319

Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., & Daniel-Vedele, F. (2002). Bay-0 x Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics*. https://doi.org/10.1007/s00122-001-0825-9

Lowry, D. B., Logan, T. L., Santuari, L., Hardtke, C. S., Richards, J. H., DeRose-Wilson, L. J., McKay, J. K., Sen, S., & Juenger, T. E. (2013). Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis. *Plant Cell*. https://doi.org/10.1105/tpc.113.115352

Nodzak, C. (2020). Introductory methods for EQTL analyses. In *Methods in Molecular Biology* (Vol. 2082, pp. 3–14). Humana Press Inc. https://doi.org/10.1007/978-1-0716-0026-9_1

Nuzhdin, S. V, Dilda, C. L., & C Mackay, T. F. (1999). *The Genetic Architecture of Selection Response: Inferences From Fine-Scale Mapping of Bristle Number Quantitative Trait Loci in Drosophila melanogaster binant isogenic (RI) chromosomes were constructed*.

Ott, J., Wang, J., & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. In *Nature Reviews Genetics* (Vol. 16, Issue 5, pp. 275–284). Nature Publishing Group. https://doi.org/10.1038/nrg3908

Serin, E. (2018). *Environmental tuning of the genetic control of seed performance: a systems genetics approach.* http://edepot.wur.nl/458202%0Ahttps://www.cabdirect.org/cabdirect/abstract/20183382153

Serin, E. A. R., Snoek, L. B., Nijveen, H., Willems, L. A. J., Jiménez-Gómez, J. M., Hilhorst, H. W. M., & Ligterink, W. (2017). Construction of a High-Density Genetic Map from RNA-Seq Data for an Arabidopsis Bay-0 × Shahdara RIL Population. *Frontiers in Genetics*, *8*(DEC), 201. https://doi.org/10.3389/fgene.2017.00201

Snoek, L. B., Terpstra, I. R., Dekter, R., Van den Ackerveken, G., & Peeters, A. J. M. (2013). Genetical genomics reveals large scale genotype-by-environment interactions in Arabidopsis thaliana. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2012.00317

Tang, W. (n.d.). *Data Integration for Potential Gene Regulation Detection Based on eQTL Analysis Data Integration for Potential Gene Regulation Detection Based on eQTL Analysis*.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., & Bork, P. (2005). STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki005

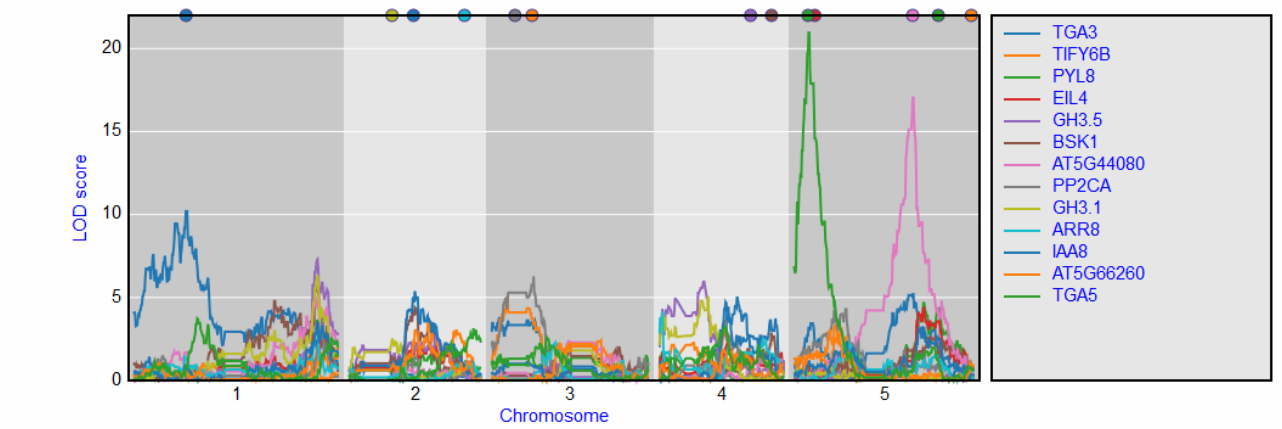
Vosman, B., Kashaninia, A., van't Westende, W., Meijer-Dekens, F., van Eekelen, H., Visser, R. G. F., de Vos, R. C. H., & Voorrips, R. E. (2019). QTL mapping of insect resistance components of Solanum galapagense. *Theoretical and Applied Genetics*. https://doi.org/10.1007/s00122-018-3239-7

Appendix I

Markers	Pathway	#Genes found	Odds ratio	Pvalue
('RSM_1_22.85', 'RSM_1_23.05')	map00902	2	11.81	0.0217
('RSM_1_22.95', 'RSM_1_23.15')	map00902	2	11.81	0.0217
('RSM_1_19.35', 'RSM_1_19.55')	map00903	2	9.1	0.0428
('RSM_1_18.15', 'RSM_1_18.35')	map03060	3	4.84	0.0431
('RSM_1_22.05', 'RSM_1_22.25')	map03022	1	inf	0.0482
('RSM_2_12.65', 'RSM_2_12.85')	map03010	83	2.36	0.0
('RSM_2_12.75', 'RSM_2_12.95')	map03010	83	2.14	0.0002
('RSM_2_12.85', 'RSM_2_13.05')	map03010	83	2.57	0.0
('RSM_2_0.25', 'RSM_2_0.45')	map00908	2	13.98	0.0206
('RSM_2_0.35', 'RSM_2_0.55')	map00908	2	13.98	0.0206
('RSM_2_0.45', 'RSM_2_0.65')	map00908	2	13.98	0.0206
('RSM_2_1.65', 'RSM_2_1.85')	map03020	2	17.86	0.0169
('RSM_2_10.35', 'RSM_2_10.55')	map00998	1	93.57	0.024
('RSM_2_13.35', 'RSM_2_13.55')	map00440	1	82.12	0.0225
('RSM_2_16.75', 'RSM_2_16.95')	map03022	1	inf	0.0482
('RSM_2_6.45', 'RSM_2_6.75')	map00750	1	inf	0.0211
('RSM_2_6.65', 'RSM_2_6.85')	map00750	1	inf	0.0211
('RSM_3_10.05', 'RSM_3_10.25')	map04120	2	inf	0.0363
('RSM_3_1.15', 'RSM_3_1.35')	map01502	1	inf	0.009
('RSM_3_10.75', 'RSM_3_10.95')	map03008	2	15.86	0.0359
('RSM_3_10.85', 'RSM_3_11.05')	map03008	2	15.86	0.0359
('RSM_3_20.35', 'RSM_3_20.55')	map00785	1	82.12	0.0225
('RSM_3_1.25', 'RSM_3_1.45')	map01502	1	inf	0.0045
('RSM_3_8.55', 'RSM_3_8.75')	map03450	1	inf	0.012
('RSM_3_8.65', 'RSM_3_8.85')	map03450	1	inf	0.012
('RSM_3_9.25', 'RSM_3_9.45')	map00750	1	49.92	0.0418
('RSM_4_14.65', 'RSM_4_14.85')	map03022	2	8.36	0.0406
('RSM_5_22.95', 'RSM_5_23.15')	map00073	2	9.61	0.0307
('RSM_5_23.05', 'RSM_5_23.25')	map00073	2	9.61	0.0307
('RSM_5_24.55', 'RSM_5_24.75')	map03008	7	3.88	0.0081
('RSM_5_15.25', 'RSM_5_15.45')	map00998	1	219.67	0.012
('RSM_5_15.35', 'RSM_5_15.55')	map00998	1	219.67	0.012
('RSM_5_9.45', 'RSM_5_9.65')	map03015	5	3.8	0.0276
('RSM_5_9.55', 'RSM_5_9.75')	map03015	5	3.41	0.0371
('RSM_5_24.65', 'RSM_5_24.85')	map00970	3	5.33	0.0307
('RSM_5_24.65', 'RSM_5_24.85')	map03008	6	4.5	0.0082
('RSM_5_25.55', 'RSM_5_25.75')	map03040	4	8.44	0.0153
('RSM_5_23.15', 'RSM_5_23.35')	map00073	2	58.09	0.0037

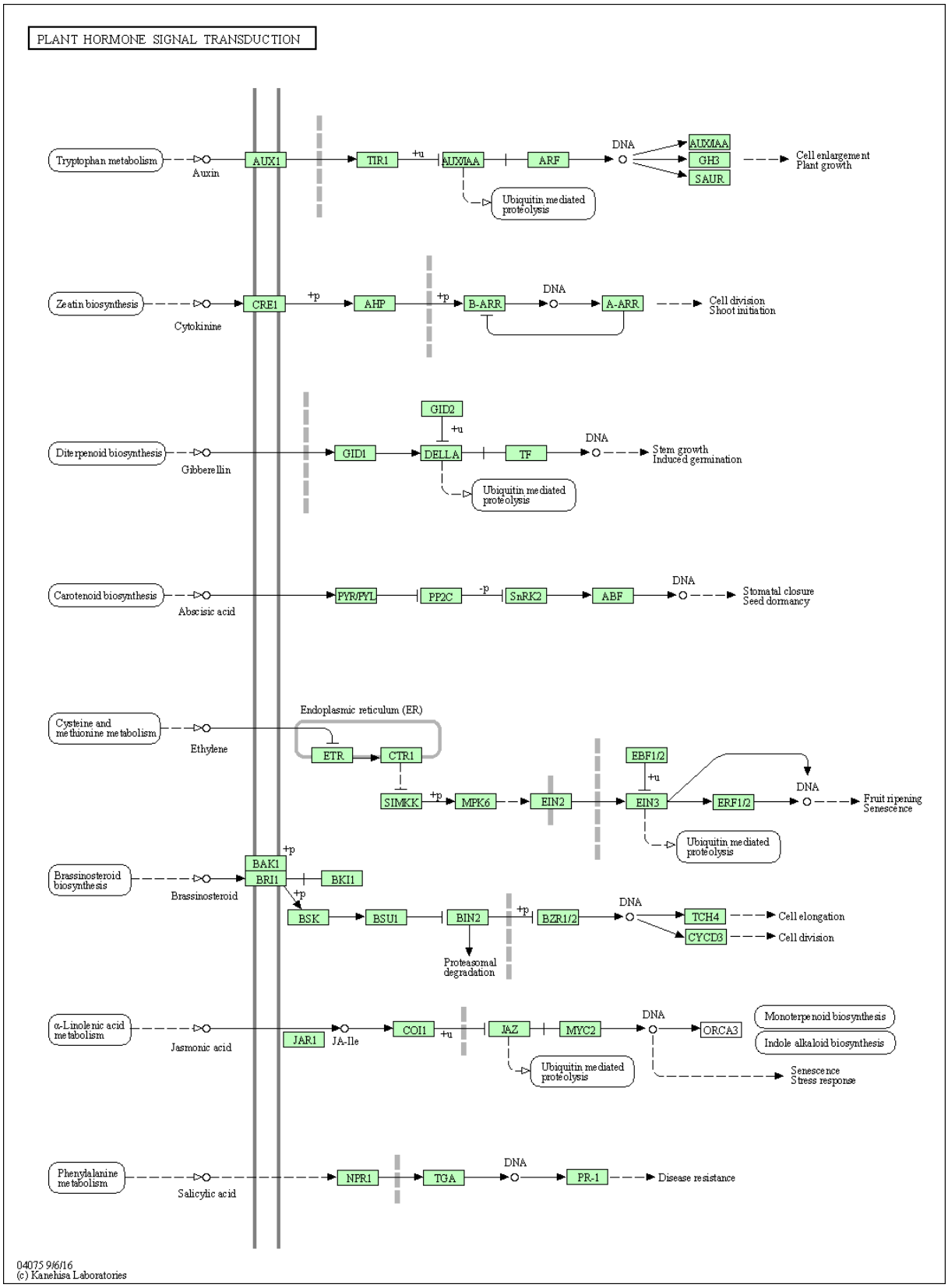
Starting marker	Ending marker	Pathway	Genes found	Total genes known in pathway	Oddratio	P-value
('c1_20384',	'c1_25698')	PANTO-PWY	6	6	inf	0.0429
('c2_04263',	'c2_06655')	PWYQT-4450	8	8	inf	0.0104
('c3_02968',	'c3_15117')	PWY-5441	4	4	inf	0.0088
('c5_06820',	'c5_08563')	PWY-3341	3	4	9.5	0.047
('c5_06820',	'c5_08563')	PROSYN-PWY	3	3	inf	0.0143
('c3_09748',	'c3_18180')	PWY-5391	6	9	6.99	0.0061
('c3_09748',	'c3_18180')	PWY-3101	6	12	3.46	0.0363
('c3_14097',	'c3_18180')	PWY-6668	2	2	inf	0.0307
('c3_08042',	'c3_12647')	map03022	15	32	2.47	0.0223
('c3_08042',	'c3_16677')	PWY-5079	6	8	11.54	0.0017
('c3_08042',	'c3_16677')	PWY-5486	6	13	3.25	0.0401
('c3_06631',	'c3_12647')	PWY-7468	2	2	inf	0.0196
('c3_06631',	'c3_12647')	PWY-6762	2	2	inf	0.0196
('c3_08042',	'c3_10996')	map03022	20	32	2.24	0.0409
('c3_09748',	'c3_16677')	PWY-7036	6	22	3.07	0.0324
('c3_09748',	'c3_12647')	CAMALEXIN-SYN	8	33	16.52	0.0
('c3_10996',	'c3_16677')	PWY-7036	10	22	2.8	0.0218
('c3_06631',	'c3_16677')	PWY-7468	2	2	inf	0.0249
('c3_06631',	'c3_16677')	PWY-6762	2	2	inf	0.0249
('c4_00641',	'c4_04877')	PWY-5143	12	12	inf	0.0419
('c1_09782',	'c1_23381')	GLUTSYNIII-PWY	4	4	inf	0.0289
('c1_11160',	'c1_24795')	PWY-5692	2	2	inf	0.0462
('c1_11160',	'c1_24795')	GLUTSYNIII-PWY	4	4	inf	0.0021
('c1_18433',	'c1_28454')	GLUTSYNIII-PWY	4	4	inf	0.0116
('c2_04263',	'c2_06655')	PWY-6066	2	2	inf	0.0462
('c3_08042',	'c3_15117')	PWYQT-4476	2	5	9.31	0.0421
('c3_06631',	'c3_16677')	PWY-5064	2	2	inf	0.0347
('c5_06820',	'c5_08563')	PWY-1186	6	7	8.66	0.0226

Appendix III



Trait ID (count=13)	Trait Name	Position	Description
AT1G22070	TGA3	1:7789133	Encodes a transcription factor. Like other TGAla-related factors%2C TGA3 has a highly conserved bZIP region and exhibits similar DNA-binding properties.
AT3G17860	JAZ3	3:6119707	JAZs are direct targets of the SCFCOI1 E3 ubiquitin-ligase and JA treatment induces their proteasome-mediated degradation. Furthermore%2C JAI3 negatively regulates the key transcriptional activator of JA responses%2C AtMYC2. The C-terminal portion of JAZ3%2C including the Jas domain%2C appears to be important for JAZ3-COI1 binding in the presence of coronatine.
AT5G53160	RCAR3	5:21560601	Encodes RCAR3%2C a regulatory component of ABA receptor. Interacts with protein phosphatase 2Cs ABI1 and ABI2. Stimulates ABA signaling.
AT5G10120		5:3169640	Ethylene insensitive 3 family protein%3B FUNCTIONS IN: sequence-specific DNA binding transcription factor activity%3B INVOLVED IN: response to karrikin%2C regulation of transcription%3B LOCATED IN: nucleus%3B EXPRESSED IN: stem%2C hypocotyl%2C root%2C flower%2C seed%3B EXPRESSED DURING: F mature embryo stage%2C petal differentiation and expansion stage%3B CONTAINS InterPro DOMAIN/s: Ethylene insensitive 3 (InterPro:IPR006957)%3B BEST Arabidopsis thaliana protein match is: Ethylene insensitive 3 family protein (TAIR:AT5G65100.1)%3B Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0%3B Bacteria - 0%3B Metazoa - 736%3B Fungi - 347%3B Plants - 385%3B Viruses - 0%3B Other Eukaryotes - 339 (source: NCBI BLink).
AT4G27260	WES1	4:13653579	encodes an IAA-amido synthase that conjugates Asp and other amino acids to auxin in vitro. Lines carrying insertions in this gene are hypersensitive to auxin. It is involved in camalexin biosynthesis via conjugating indole-3-carboxylic acid (ICA) and cysteine (Cys).
AT4G35230	BSK1	4:16755092	Encodes BR-signaling kinase 1 (BSK1)%2C one of the three homologous BR-signaling kinases (BSK1%2C AT4G35230%3B BSK2%2C AT5G46570%3B BSK3%2C AT4G00710). Mediates signal transduction from receptor kinase BRI1 by functioning as the substrate of BRI1. Plasma membrane localized.
AT5G44080		5:17737874	Basic-leucine zipper (bZIP) transcription factor family protein%3B FUNCTIONS IN: sequence-specific DNA binding transcription factor activity%3B INVOLVED IN: regulation of transcription%2C DNA-dependent%3B LOCATED IN: chloroplast%3B EXPRESSED IN: 24 plant structures%3B EXPRESSED DURING: 13 growth stages%3B CONTAINS InterPro DOMAIN/s: Basic-leucine zipper (bZIP) transcription factor (InterPro:IPR004827)%2C bZIP transcription factor%2C bZIP-1 (InterPro:IPR011616)%3B BEST Arabidopsis thaliana protein match is: G-box binding factor 4 (TAIR:AT1G03970.1)%3B Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0%3B Bacteria - 0%3B Metazoa - 736%3B Fungi - 347%3B Plants - 385%3B Viruses - 0%3B Other Eukaryotes - 339 (source: NCBI BLink).
AT3G11410	PP2CA	3:3583782	Encodes protein phosphatase 2C. Negative regulator of ABA signalling. Expressed in seeds during germination. mRNA up-regulated by drought and ABA.
AT2G14960	GH3.1	2:6451319	encodes a protein similar to IAA-amido synthases. Lines carrying an insertion in this gene are hypersensitive to auxin.
AT2G41310	RR3	2:17221840	Encodes an A- type response Regulator that is primarily expressed in the root and is involved in cytokinin-mediated signalling.
AT2G22670	IAA8	2:9636346	Encodes a transcriptional repressor of the auxin response that is auxin inducible and is involved in lateral root formation.
AT5G66260		5:26471224	SAUR-like auxin-responsive protein family %3B CONTAINS InterPro DOMAIN/s: Auxin responsive SAUR protein (InterPro:IPR003676)%3B BEST Arabidopsis thaliana protein match is: SAUR-like auxin-responsive protein family (TAIR:AT4G36110.1)%3B Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0%3B Bacteria - 0%3B Metazoa - 736%3B Fungi - 347%3B Plants - 385%3B Viruses - 0%3B Other Eukaryotes - 339 (source: NCBI BLink).
AT5G06960	OBF5	5:2154746	Encodes a basic leucine zipper (B-ZIP) containing protein that interacts with NPR1 to promote expression of salicylic acid induced genes. Binds the ocs-element.

Appendix IV


PLANT HORMONE SIGNAL TRANSDUCTION
Tryptophan metabolism
Auxin
AUX1
TIR1
+u
AUX/IAA
ARF
DNA
AUX/IAA
GH3
SAUR
Cell enlargement
Plant growth
Ubiquitin mediated proteolysis
Zeatin biosynthesis
Cytokinine
CRE1
+p
AHP
+p
B-ARR
DNA
A-ARR
Cell division
Shoot initiation
Diterpenoid biosynthesis
Gibberellin
GID2
+u
GID1
DELLA
TF
DNA
Stem growth
Induced germination
Ubiquitin mediated proteolysis
Carotenoid biosynthesis
Abscisic acid
PYR/PYL
PP2C
-p
SnRK2
ABF
DNA
Stomatal closure
Seed dormancy
Cysteine and methionine metabolism
Ethylene
Endoplasmic reticulum (ER)
ETR
CTR1
EBF1/2
+u
SIMKK
+p
MPK6
EIN2
EIN3
DNA
ERF1/2
Fruit ripening
Senescence
Ubiquitin mediated proteolysis
Brassinosteroid biosynthesis
Brassinosteroid
+p
BAK1
BRI1
BKI1
+p
BSK
BSU1
BIN2
+p
BZR1/2
DNA
TCH4
Cell elongation
CYCD3
Cell division
Proteasomal degradation
α-Linolenic acid metabolism
Jasmonic acid
JAR1
JA-Ile
COI1
+u
JAZ
MYC2
DNA
ORCA3
Monoterpenoid biosynthesis
Indole alkaloid biosynthesis
Ubiquitin mediated proteolysis
Senescence
Stress response
Phenylalanine metabolism
Salicylic acid
NPR1
TGA
DNA
PR-1
Disease resistance
04075 9/6/16
(c) Kanehisa Laboratories